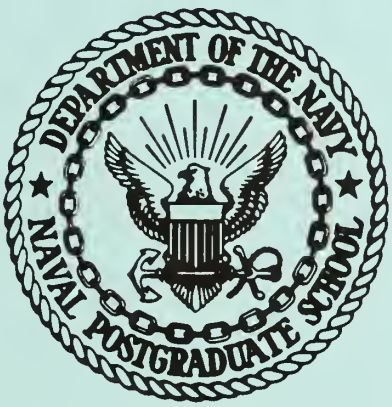
A GOODNESS OF FIT TEST FOR BIVARIATE NORMAL DISTRIBUTIONS

James Edward Miller

United States
Naval Postgraduate School



THESIS

A GOODNESS OF FIT TEST FOR
BIVARIATE NORMAL DISTRIBUTIONS

by

James Edward Miller

April 1970

A Goodness of Fit Test for Bivariate Normal Distributions

by

James Edward Miller
Lieutenant Colonel, United States Marine Corps
B.S., University of Utah, 1964

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
April 1970

ABSTRACT

This paper is an investigation of a goodness of fit test for bivariate normal distributions. The test procedure is based on random linear functions of bivariate normal random variables. The test makes use of the maximum Kolmogorov D(M) statistic over the linear functions which are computed. An estimate of the distribution of M is obtained by computer simulation. No attempt is made to determine the power of the test.

TABLE OF CONTENTS

I. INTRODUCTION ---------- 7

II. THEORY ---------- 10

III. EMPIRICAL RESULTS ---------- 18

IV. RANDOM VARIABLE GENERATION ---------- 23

V. SUMMARY AND CONCLUSIONS ---------- 27

COMPUTER PROGRAM ---------- 32

BIBLIOGRAPHY ---------- 39

INITIAL DISTRIBUTION LIST ---------- 40

FORM DD 1473 ---------- 41

LIST OF TABLES

I. Distribution (Relative Frequency) of Kolmogorov-Smirnov Statistics that Result from 100 Linear Combinations of 100 Bivariate Random Vectors from Two Bivariate Normal Distributions ---------- 29

II. Distributions (Relative Frequency) of 500 Maximum Kolmogorov-Smirnov Statistics, Each of which was Derived from 25 Linear Combinations of Samples from Various Bivariate Normal Distributions -------- 30

III. Distribution (Relative Frequency) of 500 Maximum Kolmogorov-Smirnov Statistics for Samples Drawn from Various Bivariate Normal Distributions with Identical Correlation Coefficient, ρ ---------- 31

I. INTRODUCTION

An important application of Statistics is to attempt to find a specific probability distribution which fits an observed sample of data. A well fitting distribution can then be used to predict values of future occurrences, relative frequencies of future occurrences, etc. There are numerous methods available for testing the goodness of fit of data in scalar form to a hypothesized univariate probability distribution. Among these are the Chi-square and the Kolmogorov-Smirnov tests, of which the Kolmogorov-Smirnov test is considered the more powerful [1]. Furthermore the Kolmogorov-Smirnov test exhibits the very attractive characteristic that it is based on a statistic which has a distribution of the random variable being sampled.

However, there appears to be a lack of methods for testing the fit of hypothesized multivariate distributions to multivariate data (observations in vector form). Furthermore, no statistic which has the desirable characteristic of being distribution free and which can be used in multivariate goodness of fit tests has been found. In fact, no such statistic may even exist. For instance, Simpson [2] has shown an example of continuous bivariate distributions for which the analog of the Kolmogorov-Smirnov statistic is dependent on the underlying distribution.

Rosenblatt [3] discusses a possible test which involves a transformation of an absolutely continuous k-variate distribution into the uniform distribution on the k-dimensional hypercube. The transformation is uniquely determined by the theoretical distribution against which the sample is to be tested. Then the transformed sample may be tested against the uniform distribution in k-dimensions. There are several disadvantages to this procedure, however. For example, the results are influenced by the manner in which the components of the observed vectors are ordered.

The purpose of this paper is to describe a goodness of fit test for testing a bivariate normal distribution (with given mean and covariance matrix) against samples of bi-component data. The test results in acceptance or rejection of the hypothesis that $F_N = F$, where F_N is the cumulative distribution function of the population of the bivariate sample vectors and F is the hypothesized cumulative distribution function. The notation in this paper closely follows the notation used by Anderson [4]. It is expected that the test developed here for the bivariate normal distribution can be extended to the case of the k-variate normal distribution. This paper is restricted to a consideration of testing the fit of samples to a distribution which has zero mean. This restriction causes no loss of generality since any distribution with a finite mean can be translated to mean zero by a linear transformation.

The goodness of fit test was developed according to the following procedures:

1) a characterization of the bivariate normal distribution is used to develop a test statistic M for use in a goodness of fit test,

2) the distributional properties of M are investigated by computer simulation.

II. THEORY

Since there appeared to be no widely known statistic for a reasonable goodness of fit test for multivariate distributions in general, and the multivariate normal distribution in particular, it seemed plausible that a statistic suitable for a goodness of fit test might be found by considering characterizations of the multivariate normal distribution.

One property which characterizes a multivariate normal distribution is given in the following theorem [4]:

> Theorem 1. A p-dimensional random variable X has a p-variate normal distribution, if an only if every linear function of X has a univariate normal distribution.

The parameters of the univariate normal distribution can be computed according to theorem 2.

> Theorem 2. Let X (a column vector with p components) be distributed according to $N_p(\mu,\Sigma)$ a multivariate normal distribution with mean (vector) μ and covariance (matrix) Σ, and let C be a row vector of p constants. Then
>
> $$Y = CX$$
>
> is distributed as univariate normal with mean $C\mu$ and variance $C\Sigma C'$ (C' is the transpose of C).

(NOTE: CX can be described as a linear combination of the components of X.)

From the characterization of the multivariate normal distribution given in theorem 1, it was felt that a suitable goodness of fit test procedure might be to test the result of a linear combination of the sample vectors (whose distribution had been hypothesized as a specific multivariate normal distribution) against the hypothesized theoretical univariate normal distribution which has been computed for the particular linear combination. Thus the problem is reduced to the univariate level and use can be made of well known univariate statistics which provide acceptable goodness of fit tests.

However, theorem 1 states that <u>every</u> linear combination of multivariate normal random variables must be univariate normal. Obviously one linear combination will not suffice for a reasonable test. It is not difficult to envision that there exists some linear function of nearly any vector sample which will transform that vector sample into one which is accepted as univariate normal. In fact, if the marginal distribution of the components are univariate normal, but the joint distribution is not multivariate normal, the linear combination consisting of one component (e.g. $Y = X_1 + 0X_2 \ldots + 0X_N = X_1$) is univariate normal. Thus a test which uses only one linear combination might be manipulated by the tester to give any results he desires.

On the other hand, it is clearly impossible to compute every linear combination of a sample. As a compromise, it was felt that a number (to be determined) of randomly selected linear combinations would serve as a representative sample upon which an overall test statistic might be based. To produce random linear combinations, the (column) sample vectors were multiplied by a (row) vector of random constants. The random components of the 'multiplying vectors' were drawn from the uniform (0,1) distribution. A uniform (0,1) distribution for the random multipliers was used because:

1) Up to multiplicative constants, essentially any linear combination of the components of the multivariate vector could be produced using coefficients from the uniform (0,1) distribution, and

2) A component of a random multiplier was equally likely to be contained in any one interval in (0,1) as in any other interval, provided the intervals were of equal length. Thus there should be no specific interval containing a 'concentration' of the multipliers which might adversely influence the performance of the goodness of fit test.

NOTE: The results of the goodness of fit test described in this paper using random multipliers from a uniform (0,1) distribution were the same as results obtained using random multipliers from a uniform (-2,2) distribution.

The Kolmogorov-Smirnov test was employed to determine acceptance (or rejection) of the hypothesis that the linear combinations of bivariate sample vectors are from the (computed) theoretical univariate distributions noted in theorem 2. As noted previously, the Kolmogorov-Smirnov test is considered more powerful than the Chi-square test. Of course, the distribution free characteristic of the Kolmogorov D statistic applies in particular to linear combinations of the components of multivariate normal random variables. A description of the Kolmogorov-Smirnov test is presented in the following paragraphs.

One method of testing the simple hypothesis, H: $F_N = F$, where F_N is the cumulative distribution of the population sampled and F is the theoretical continuous distribution proposed for the population, is the Kolmogorov D statistic [5]. The asymptotic distribution of D was investigated by Kolmogorov and tabulated by Smirnov [6] and, for small sample sizes, by Massey [7].

The Kolmogorov D statistic is derived from the sample cumulative distribution function, S_N, and the proposed theoretical cumulative distribution function, F, as follows; Let $Y_1,\ldots Y_N$ be a random sample from a continuous population with cumulative distribution function F. Let $Z_1,\ldots Z_N$ be the ordered statistics of Y, so that

$$-\infty < Z_1 \leq Z_2 \leq \ldots \leq Z_N < \infty$$

The sample cumulative distribution, then, is

$$S_N(x) = \begin{cases} 0 & x < Z_1 \\ j/N & Z_j \leq x < Z_{j+1}, \quad j = 1,\ldots,N-1 \\ 1 & x \geq Z_N \end{cases}$$

The Kolmogorov D statistic is defined as

$$D = \sup_x |S_N(x) - F(x)| .$$

and can be described roughly as the maximum deviation of the sample cumulative distribution function from the proposed theoretical cumulative distribution function. The D statistic has the property that its distribution does not depend upon the underlying distribution F. Clearly, it is dependent on the sample size N, because the sample cumulative distribution functions, S_N, takes as values only multiples of 1/N. Naturally the D statistic approaches zero, almost surely, as N becomes large without limit, providing the sample is actually from a population with distribution F. Critical values, T, of the D statistic are obtained from the tabulated distributions and are used with a sample to determine acceptance or rejection (reject if $D \geq T$) of the hypothesis that $F_N = F$.

One value of the Kolmogorov D statistic is derived from each linear combination of a bivariate sample of vectors. For example, let $\overline{X} = (X_1,\ldots X_N)$ where $X_i = (x_{i1},x_{i2})$ and x_{ij} is a scalar, be a random sample of size N of bivariate

random vectors. A linear combination $\overline{Y} = C\overline{X}$, where $C = (c_1, c_2)$ and c_i is a scalar, is a vector of N scalars, $(Y_1, \ldots Y_N)$. Let Σ be the hypothesized covariance matrix of the distribution of $\overline{X}$. To obtain a rough test of the hypothesis that the distribution of $\overline{X}$ is bivariate normal with covariance matrix Σ (and mean zero), one may test the hypothesis that $\overline{Y} = C\overline{X}$ is distributed as univariate normal with variance $C\Sigma C'$ (and mean zero). A Kolmogorov-Smirnov test to determine the acceptance of the hypothesis when one particular value of C, say C_i, is used to compute a linear combination will yield one value of D, that is

$$D = \sup_y \left| F_{\overline{Y}}(y) - S_N^*(y) \right|$$

where $F_{\overline{Y}}$ is the hypothesized univariate normal cumulative distribution of $\overline{Y}$ and S_N^* is the sample cumulative distribution of the transformed sample. When the procedure listed above is repeated for a different value of C, say C_j, then another value of D, which may or may not be identical to the first value, is obtained. If every value of D obtained with various values of C is less than the critical value of D for the given sample size and level, then the hypothesis is accepted. Likewise if every value of D obtained from using various values of C is greater than the critical value, then the hypothesis is rejected. However, some linear combinations of typical samples from a bivariate normal population can be expected to give values of D

exceeding the critical value while other linear combinations may give values of D less than the critical value. To eliminate the ambiguity of such results, another statistic must be used, preferably one whose distribution function can be readily tabulated or computed. For this purpose we use the maximum of the D statistics, M, derived from Kolmogorov-Smirnov tests of a large number, m, of linear combinations of the sample. That is,

$$M = \max_i \sup_y \left| S_{C_i\overline{X}}(y) - F_i^*(y) \right| \quad i = 1,2,\ldots m$$

where $S_{C_i\overline{X}}$ is the sample cumulative distribution function of the linear combination $C_i\overline{X}$ and $F_i(y)$ is the hypothesized theoretical cumulative distribution of the linear combination $C_i\overline{X}$.

The values of the D statistics derived from linear combinations of a particular sample appear to be more highly dependent on the sample than on the random multiplying vectors (see Section III, Empirical Results). Therefore the maximum D might be expected to be a result of only the sample so that rejection or acceptance of the hypothesis H: $F_N = F$, where F_N is the cumulative distribution function of the population sampled, and F is the proposed theoretical cumulative distribution function, would depend only on the bivariate sample.

The distribution of M, as defined above, appeared to be intractable to get in closed mathematical form. However,

the properties of the distribution of M were investigated for several cases by examining empirical data obtained from computer simulation. The data was produced by generating samples of a specified bivariate normal distribution, computing random linear combinations of the sample vectors, and recording the maximum of the resulting Kolmogorov D statistics. This procedure was repeated to give several lists of 500 values of M. Each list of 500 values of M was derived from bivariate sample vectors with different underlying bivariate normal distributions. The empirical results and generating techniques are discussed in Sections III and IV.

III. EMPIRICAL RESULTS

In order to obtain the empirical data to study the distribution of M, the maximum D statistic, a computer program was written to accomplish the following for specific selections of the covariance matrix, Σ:

1. Generate a sample of desired size of bivariate normal random vectors from the given distribution,

2. Generate the desired number of multiplying vectors, each of which would produce one linear combination of the sample vectors,

3. Compute the linear combinations of the sample vectors by vector multiplication,

4. Perform a Kolmogorov-Smirnov test on each univariate sample obtained as a result of a linear combination and record the resulting value of D. The values of

$$D = \sup_x |S_N(x) - F(x)|$$

when S_N and F are functions previously defined, were determined at values of .01K (K = 1,...100) for the proposed theoretical distribution F. For example, the value of the sample cumulative distribution, $S_N(x)$, was evaluated at each point x_i where $F(x_i)$ was multiple of .01, D being assumed to be the maximum value of the 100 differences

$$|S_N(x_i) - F(x_i)|$$

5. Record the maximum value, over the linear combinations performed on each sample, of the D statistics produced by each particular sample.

The initial simulation procedure generated sets of 100 random vectors from a bivariate normal distribution with covariance matrix Σ. One hundred 'random' linear combinations of each set of vectors were computed and the D Statistic derived from each linear combination was recorded. The results of this simulation indicated that the D statistics for a given sample were grouped within an interval approximately .05 units in length, but the location of the interval was dependent upon the particular sample. This phenomenon suggested that the value of D is dependent on the sample to a higher degree than it is on the linear function used. The results of five such simulations for samples from each of two different bivariate normal distributions are summarized in Table I. Note that with sample number 1b, several D values exceeded the univariate Kolmogorov-Smirnov critical value at the .05 level of significance. Thus, using the univariate Kolmogorov-Smirnov critical value, the hypothesis that the sample was from the underlying distribution from which it was generated would have been rejected for some linear combinations and accepted for others. But using a critical value (determined by level of significance and sample size) for maximum D would have eliminated the ambiguity.

It was also found that the relative frequency, within each interval of length .01, of the D statistics remained nearly constant as the number of linear combinations was varied.

The data produced by the simulation procedure described above led to the consideration of using the maximum D Statistic for testing a bivariate sample against the proposed bivariate normal distribution. This simulation data indicates that the maximum D statistic derived from random linear combinations of samples from a bivariate normal population had the desirable characteristics that:

1) A unique maximum is obtained for each sample, independent of the random multipliers, provided a sufficient number of linear combinations are computed, and

2) The maximum value is obtained from various linear combinations, at least one of which could be randomly selected, with high probability, in as few as 25 trials (selections) of multiplying vectors. In all cases investigated, including those listed in Table I, the same value of M was achieved over 25 linear combinations as was achieved in 100 linear combinations for each particular sample.

As noted in Section II, the exact distribution of M was found to be intractable. Therefore, in order to study some of the characteristics of the distribution of M, another computer simulation procedure was used to produce a large sample of M. The simulation procedure may be described as follows:

1) A sample of 25 vectors was generated from a predetermined bivariate normal distribution. Since D, and therefore M, are dependent on sample size, it was recognized that data obtained by this simulation would pertain

to samples of size 25 only. However, one might expect the characteristics of the distribution of M to be similar for all sample sizes.

2) Twenty-five randomly selected multiplying vectors were generated so that 25 linear combinations of each sample were produced. The maximum D over the resulting 25 univariate samples was recorded. (From the initial simulation, it was expected that 25 linear combinations would produce the maximum D for any sample.) A total of 500 M statistics, all derived from the same underlying bivariate normal distribution, were produced.

The simulation procedures were repeated for different parameter values of the underlying bivariate normal distribution to produce five sets of 500 statistics. Thus, each set of 500 values was derived from linear combinations of samples drawn from a different bivariate normal distribution.

The results of the simulation described above are summarized in Table II. Unfortunately, it appears that there is not a simple statistical relationship between the distributions of the M statistics obtained with the samples drawn from different bivariate distributions. And, of course, if each different underlying distribution (of the sample) produces a different distribution of the M statistic, it would be impossible to tabulate values of all distribution functions of M.

It did not seem unlikely that a similarity or other relationship existed between the distributions of M statistics obtained from samples which were derived from bivariate normal distributions with identical correlation coefficients. Therefore, a final simulation procedure was repeated, using samples drawn from several non-identical bivariate normal distributions with constant correlation coefficients. The results are summarized in Table III.

Although there is a notable similarity between the distributions of M statistics derived from bivariate normal distributions with identical correlation coefficients (ρ), the hypothesis that the distributions are identical was rejected by a Kolmogorov-Smirnov test at the .05 level of significance. This is also readily apparent for the case in which $\rho = .3162$. Note that the difference in the means of the samples of M is .0059, whereas one standard deviation of the mean (computed from the sample standard deviation) is approximately .0022. Thus the means are nearly three standard deviations apart, which suggests that the distributions are not the same.

An interpretation of these results and how they may be applied to a possible goodness-of-fit test for the bivariate normal distribution is discussed in Section V, Summary and Conclusions.

IV. RANDOM VARIABLE GENERATION TECHNIQUES

In order to study the distribution of the M statistics described in Sections II and III, it was necessary to produce a large number of random variables from various bivariate normal distributions. There are several possible methods which might be used to generate the bivariate normal random variables on a computer. One method would be to generate independent normal random variables and perform an appropriate transformation on them which will produce a bivariate normal random vector. For example, to generate random vectors from a bivariate normal distribution with mean zero and covariance matrix Σ, where Σ is symmetric and positive definite, one could use the following procedure.

1) Generate two independent random variables, from a normal (0,1) distribution, so that $X = (X_1, X_2)$ is bivariate normal (0,I) where I is the identity matrix.

2) Perform the transformation $Z = CX$, where C satisfies $CC' = \Sigma$. Then $Z = (Z_1, Z_2)$ is bivariate normal $(0,\Sigma)$.

In this study the bivariate normal random vectors were generated using a conditional distribution approach. It is a well known characteristic of the bivariate normal distribution that if $X = (X_1, X_2)$ is distributed bivariate normal (μ,Σ) where $\mu=(\mu_1,\mu_2)$ and

$$\Sigma = \begin{pmatrix} \sigma_{12} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

then X_2 is distributed univariate normal (μ_2,σ_{22}). It is also well known that the conditional distribution of X_1, given $X_2 = X_2$ is univariate normal $[\mu_1+\sigma_{12}\sigma_{22}^{-1}(X_2-\mu_2),\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21}]$. Therefore, after generating X_2 from a univariate normal (μ_2,σ_{22}) distribution, the conditional distribution of X_1, given $X_2 = x_2$, was computed and X_1 was then generated from that univariate normal distribution.

To verify that this produces a random vector with the characteristics of the given bivariate normal distribution, consider the following:

Let $\mu = (0,0) = (\mu_1,\mu_2)$

Generate $X_2 = x_2$ from its marginal distribution, $N(0,\sigma_{22})$. Then,

$$E(X_2) = 0$$

and

$$E[(X_2 - \mu_2)^2] = E[(X_2)^2] = \sigma_{22}$$

Generate V, independent of X_2, from univariate normal (0,1) distribution.

Now

$$E(V) = 0, \text{ and}$$

$$E(V^2) = 1$$

Now let $X_1 = (\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})^{\frac{1}{2}} V + \sigma_{12}\sigma_{22}^{-1}(X_2)$.

X_1 is univariate normal $(\sigma_{12}\sigma_{22}^{-1}(X_2), \sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})$, and

$$E(X_1) = E[(\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})\ V + \sigma_{12}\sigma_{22}^{-1}(X_2)] = 0 = \mu_1$$

Similarly, the covariance between X_1 and X_2 is

$$\begin{aligned} Cov(X_1,X_2) &= E[(X_1-\mu_1)(X_2-\mu_2)] = E(X_1X_2) \\ &= E\{[(\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})^{\frac{1}{2}} V + \sigma_{12}\sigma_{22}^{-1}(X_2)]X_2\} \\ &= (\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})^{\frac{1}{2}} E(V\cdot X_2)+\sigma_{12}\sigma_{22}^{-1}\ E(X_2). \end{aligned}$$

Since V and X_2 are independent,

$$E(V \cdot X_2) = E(V)E(X_2) = 0.$$

Continuing from above,

$$Cov(X_1,X_2) = 0 + \sigma_{12}\sigma_{22}^{-1}(\sigma_{22}) = \sigma_{12} = \sigma_{21}\ .$$

The variance of X_1 is

$$\begin{aligned} E[(X_1 - \mu_1)^2] &= E(X_1^2) \\ &= E\{[(\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})^{\frac{1}{2}} V + \sigma_{12}\sigma_{22}^{-1}(X_2)]^2\} \\ &= (\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})E(V^2)+(\sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21})^{\frac{1}{2}}E(V) \\ &\quad (\sigma_{12}\sigma_{22}^{-1})\ E(X_2) + (\sigma_{12}\sigma_{22}^{-1})^2\ E(X_2^2) \\ &= \sigma_{11}-\sigma_{12}\sigma_{22}^{-1}\sigma_{21} + \sigma_{12}\sigma_{22}^{-1}\sigma_{12} = \sigma_{11}, \end{aligned}$$

since $\sigma_{12} = \sigma_{21}$.

Incidentally, this technique can be extended to a method of generation of p-component multivariate normal random variables, since the distribution of $X_i (i=1,2,...\rho)$ given any $X_j = X_j (j=1,2,...\rho, j \neq i)$, is also a normal distribution whose parameters may be computed.

Standard computer routines were used to generate the univariate normal and uniform random variables required for the simulation procedure previously described. The routines are shown in the computer program under Subroutine RANDU, (for uniform random variables) and Subroutine GAUSS (for univariate normal random variables).

V. SUMMARY AND CONCLUSIONS

The empirical data indicates that the distribution of the maximum of the D statistics (M), derived from Kolmogorov-Smirnov tests of linear combinations of samples from bivariate normal distributions, was dependent upon the covariance matrix of the underlying distribution of the sample. Therefore it would be impossible to tabulate the distribution of M except for specific parameters of the underlying distribution.

However, a goodness of fit test for the bivariate normal distribution can be constructed using the M statistic. The test might consist of using a simulation procedure, similar to that used in this paper, to produce a sample distribution of M. This distribution of M would be derived from samples which are from a bivariate distribution identical to the proposed hypothesized bivariate distribution. Then a critical value of M, for a test with level of significance $\alpha = \alpha_o$, may be established as the value at which the $(1 - \alpha_o)$ percentile point of the distribution of M occurs. Obviously the number of linear combinations and the number of M statistics for development of the sample distribution of M must be determined by the experimentor performing the test. (Note that the size of the samples generated in the simulation procedure must be identical to the size of the sample to be tested.

For example, suppose one wishes to test the hypothesis that a sample of N vectors was drawn from a population whose distribution is bivariate normal $(0,\Sigma)$ where $\Sigma = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$, (one of the distributions for which M is tabulated in Table II). If N = 25, then the distribution of the M statistics is shown in Table II, listed under the appropriate covariance matrix. For $\alpha = .05$, the critical value of M is .33, the value at which the .95 percentile point of the distribution occurs. Then 25 random linear combinations of the sample vectors would be computed and the 25 resulting univariate samples tested against the computed univariate distribution by a Kolmogorov-Smirnov test. If the maximum of the 25 D statistics thus obtained is greater than .33, the hypothesis is rejected. Otherwise the hypothesis is accepted.

There are obviously many interesting aspects concerning this (and other) multivariate goodness of fit tests which should be investigated. For example the power of the test described in this paper, when applied to samples from distributions other than the bivariate normal, might be investigated. Also, a goodness of fit test based on a statistic other than M (e.g., the mean or variance of D obtained from linear combinations of the sample components) might prove to be interesting. It is, of course, desirable to find a "reasonable" statistic for which the distribution may be found and tabulated.

TABLE I

Distribution (Relative Frequency) of Kolmogorov-Smirnov Statistics (D_N) that Result from 100 Linear Combinations of 100 Bivariate Random Vectors from Two Bivariate Normal Distributions

Value of K-S Statistic	$\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ SAMPLE NUMBER					$\Sigma = \begin{bmatrix} 5 & 2 \\ 1 & 2 \end{bmatrix}$				
	1a	1b	1c	1d	1e	2a	2b	2c	2d	2e
.04								2		
.05	4			21	25		10	38		
.06	11			50	66		35	41		
.07	26		2	24	9		23	19	8	
.08	6		22	5		17	9		28	
.09	44	20	46			22	14		36	6
.10	9	17	30			32	9		18	31
.11		4				29			10	29
.12		9								20
.13		39								14
.14		11								
.15										

NOTE: 1) The critical value of the univariate Kolmogorov-Smirnov statistic for this sample size and $\alpha = .05$, is .136.

2) Σ = Covariance matrix of the distribution from which the sample was drawn.

TABLE II

Distribution (Relative Frequency) of 500 Maximum Kolmogorov-Smirnov (M) Statistics Each of which was Derived from 25 Linear Combinations of Samples from Various Bivariate Normal Distributions

Range of Max. K-Statistic (M)	Covariance Matrix of Distribution of Population from which Samples were Drawn				
	$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$	$\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & -1.8 \\ -1.8 & 1 \end{bmatrix}$	$\begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$
.06-.07		1			
.07-.08		3			
.08-.09		7			
.09-.10		22			
.10-.11		22			1
.11-.12	5	40			4
.12-.13	8	44	5		6
.13-.14	21	26	7	5	9
.14-.15	31	55	18	12	21
.15-.16	28	34	19	19	32
.16-.17	38	37	31	23	43
.17-.18	34	32	24	30	59
.18-.19	41	22	41	35	35
.19-.20	42	31	42	42	38
.20-.21	39	20	50	40	31
.21-.22	30	25	41	36	26
.22-.23	27	16	33	32	31
.23-.24	28	14	34	49	21
.24-.25	25	12	31	38	33
.25-.26	29	8	30	23	21
.26-.27	20	6	20	21	20
.27-.28	8	5	15	21	14
.28-.29	18	6	15	21	15
.29-.30	11	4	10	13	15
.30-.31	7	3	7	8	8
.31-.32	8	2	7	9	7
.32-.33	0	2	7	8	7
.33-.34	6	1	7	10	11
.34-.35	2		1	2	3
.35-.36	3		3	3	5
.36-.37	2		2		2
.37-.38					1
Mean*	.2033	.1624	.2151	.2237	.2061
Variance*	.0028	.0027	.0024	.0027	.0028

NOTE: *Sample mean and Sample variance of M statistics.

2) Each sample size was 25 bivariate vectors.

TABLE III

Distribution (Relative Frequency) of 500 Maximum Kolmogorov-Smirnov Statistics (M) for Samples Drawn from Various Bivariate Normal Distributions with Identical Correlation Coefficient, ρ

Corr. Coeff. (ρ)	$\rho = .6$		$\rho = 0$		$\rho = \frac{\sqrt{10}}{10} = .3162$	
Σ / Range of M	$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$	$\begin{bmatrix} 1 & .6 \\ .2 & 1 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 1 & .3162 \\ .3162 & 1 \end{bmatrix}$
.10-.11		1			1	
.12	5	3			4	
.13	8	8	2	5	6	4
.14	21	18	3	7	9	8
.15	31	27	19	12	21	21
.16	28	51	31	31	32	29
.17	38	41	26	33	43	31
.18	34	38	25	46	59	26
.19	41	51	39	55	35	44
.20	42	31	35	29	38	39
.21	39	26	38	37	31	41
.22	30	28	43	34	26	35
.23	27	27	45	31	31	33
.24	28	21	37	23	21	44
.25	25	26	23	24	33	15
.26	29	16	30	23	21	28
.27	28	25	22	26	20	16
.28	8	9	10	19	14	17
.29	18	16	21	13	15	18
.30	11	6	14	16	15	11
.31	7	9	11	7	8	9
.32	8	5	12	9	7	9
.33	0	3	4	3	7	3
.34	6	1	1	2	11	4
.35	2	2	4	6	3	3
.36	3	5	4	6	5	3
.37	1	3	2	3	2	2
.38					1	
Mean*	.2032	.1998	.2168	.2130	.2061	.2120
Variance*	.0028	.0028	.0026	.0027	.0028	.0026

NOTE: 1) *Mean and variable of M.

2) Sample size - 25 vectors.

3) Σ = covariance matrix or distribution of population from which samples were drawn.

COMPUTER PROGRAM

```
C
C
C     THIS PROGRAM INVESTIGATES THE DISTRIBUTION OF M FOR SAMPLES
C     FROM A BIVARIATE NORMAL DISTRIBUTION
C
C
      DIMENSION SIGMA (2,2)
C
C     SIGMA IS THE COVARIANCE MATRIX FOR THE BIVARIATE NORMAL
C     RANDOM VECTORS
C
      DATA SIGMA/5.0,3.0,3.0,5.0/
C
C     THIS DATA STATEMENT ENTERS THE VALUES OF THE COVARIANCE MATRIX
C
      DIMENSION TSTAT (500)
C
C     TSTAT IS THE LIST OF M STATISTICS FROM EACH SAMPLE
C     THE DIMENSION OF 500 IS ARBITRARILY ASSIGNED
C
      DIMENSION RANVEC (100,2)
C
C     RANVEC IS THE LIST OF RANDOM VECTORS GENERATED.
C     THE DIMENSION OF 100 WAS ARBITRARILY ASSIGNED
C
      DIMENSION RDMLT (100,2)
C
C     RDMLT IS THE LIST OF RANDOM U(0,1) MULTIPLIERS GENERATED.
C     THE DIMENSION OF 100 WAS ARBITRARILY ASSIGNED
C
      DIMENSION XMAT (100)
C
C     XMAT IS THE LIST OF VARIABLES PRODUCED BY THE MULTIPLICATION OF
```

```
C          RANDOM VECTORS BY THE RANDOM MULTIPLIERS.
C          THE DIMENSION OF XMAT SHOULD BE THE SAME AS RANVEC
C
      DIMENSION RNORV (500)
C
C          RNORV IS THE RESULT OF TRANSFORMING THE (UNIVARIATE) VARIABLES
C          OF XMAT TO NORMAL (0,1)
C
      COMMON RNORV
C
      NVEC = 100
C          NVEC IS THE NUMBER OF RANDOM VECTORS TO BE GENERATED PER SAMPLE
C          NVEC MUST BE INITIALIZED PRIOR TO EXECUTION OF THE PROGRAM
C
      NMLT = 100
C          NMLT IS THE NUMBER OF RANDOM MULTIPLIERS TO BE GENERATED
C          NMLT MUST BE INITIALIZED PRIOR TO EXECUTION OF THE PROGRAM
C
      NSTAT = 500
C
C          NSTAT IS THE NUMBER OF VALUES OF M USED TO INVESTIGATE THE
C          DISTRIBUTION OF M
C          NSTAT MUST BE INITIALIZED PRIOR TO EXECUTION OF THE PROGRAM
C
      IX = 7345189
      WRITE (6, 977) NVEC
  977 FORMAT (1H1, T20, 'THE SAMPLE SIZE FOR THIS PROGRAM IS ',I5,
     1' VECTORS')
      WRITE (6, 604) NMLT
  604 FORMAT(//T20,'THE NO. OF LINEAR COMBINATIONS (MULTIPLIERS) FOR THI
     1S SAMPLE IS ',I5)
C
      WRITE(6,820)SIGMA(1,1),SIGMA(1,2),SIGMA(2,1),SIGMA(2,2)
  820 FORMAT(//T20,'THE COVARIANCE MATRIX FOR THIS SAMPLE IS ',4F10.4//)
C
      DO 675 JM = 1, NSTAT
C
C
C          THIS PORTION OF THE PROGRAM GENERATES THE BIVARIATE NORMAL
C          RANDOM VECTORS
C
      Z1 = SQRT (SIGMA (2,2))
      AM1 = 0.0
      DO 100 I = 1, NVEC
      CALL GAUSS (IX, Z1, AM1, V)
```

```
      RANVEC (I,2) = V
      AM2 = SIGMA(1,2) * (1.0 / SIGMA(2,2)) * RANVEC(I,2)
      IF (I .GT. 1) GO TO 171
      S2 = SIGMA(1,1) - SIGMA(1,2) * (1.0 / SIGMA(2,2)) * SIGMA(2,1)
      S1 = SQRT (S2)
  171 CALL GAUSS (IX, S1, AM2, V2)
      RANVEC (I,1) = V2
  100 CONTINUE
C
CCCCC
CCCCC    THIS PORTION OF THE PROGRAM GENERATES THE RANDOM MULTIPLIERS
CCCCC    WHICH PRODUCE LINEAR COMBINATIONS
CCCCC
C
      DO 103 I = 1, NMLT
      CALL RANDU (IX, IY, XFL)
      IX = IY
      RDMLT (I,1) = XFL
      CALL RANDU (IX, IY, ZFL)
      IX = IY
      RDMLT (I,2) = ZFL
  103 CONTINUE
C
CCCCC
CCCCC    THIS PORTION OF THE PROGRAM MULTIPLIES THE RANDOM VECTORS BY THE
CCCCC    MULTIPLIERS AND TRANSFORMS THE RESULT TO NORMAL(0,1) VARIABLES
CCCCC
      TTMAX = 0.0
C
      DO 140 J = 1, NMLT
      DO 104 I = 1, NVEC
      XMAT (I) = RANVEC(I,1) * RDMLT(J,1) + RANVEC(I,2) * RDMLT(J,2)
      IF (I .GT. 1) GO TO 1039
      VRNCE = SIGMA(1,1) * (RDMLT(J,1)**2) + (2 * SIGMA(1,2) * RDMLT(J,1
     1) * RDMLT(J,2)) + SIGMA(2,2) * RDMLT(J,2)**2
 1039 RNORV (I) = XMAT (I) / SQRT (VRNCE)
  104 CONTINUE
C
CCCCC
CCCCC    THIS PORTION OF THE PROGRAM PERFORMS A K-S TEST ON THE LIST OF
CCCCC    NORMAL (0,1) VARIABLES
CCCCC
CCCCC    THE MAXIMUM KOLMOGOROV STATISTIC FROM EACH TEST IS RETURNED
CCCCC
      CALL KSTEST (KTST, TMAX, NVEC)
C
CCC      TMAX IS THE MAXIMUM VALUE OF THE K-S STATISTIC FROM EACH TEST
C
```

```
C          NVAR IS NUMBER OF VARIABLES IN THE SAMPLE
C
C          IF KNORV IS ACCEPTED AS NORMAL(0,1), KTST EQUALS ONE
C          IF RNORV IS REJECTED AS NORMAL(0,1), KTST EQUALS ZERO
C
      IF (TMAX .GE. TTMAX) TTMAX = TMAX
  140 CONTINUE
      TSTAT(JM) = TTMAX
  670 CONTINUE
C
C
C          THIS PORTION OF THE PROGRAM COMPUTES THE MEAN, VARIANCE AND
C          RELATIVE FREQUENCY OF THE M STATISTICS
C
      DO 678 JA = 1, 100
      RSUM = JA / 100.0
      KSUM = 0
      DO 677 JFK = 1, NSTAT
      NSUM = 0
      IF (TSTAT(JFK) .GE. RSUM) NSUM = 1
      KSUM = KSUM + NSUM
  677 CONTINUE
      WRITE (6, 679) RSUM, KSUM
  679 FORMAT(/T20,'THE NUMBER OF MAX. K-S STATISTICS .GE. ',F5.4,' IS ',
     1I4)
  678 CONTINUE
      SUM = 0.0
      DO 690 LCL = 1, 100
  690 SUM = SUM + TSTAT(LCL)
      RBAR = SUM / (1.0 * NSTAT)
      WRITE (6, 691) RBAR
  691 FORMAT (/T20,'THE MEAN OF MAX. K-S STATISTICS IS ',F7.5)
      SUM = 0.0
      DO 692 LC = 1, 100
  692 SUM = SUM + ((TSTAT(LC) - RBAR)**2)
      VBAR = SUM / (1.0 * NSTAT)
      STDEV = SQRT(VBAR)
      WRITE (6, 693) VBAR
  693 FORMAT (/T20,'THE VARIANCE OF THE MAX. K-S STATISTICS IS ',F10.7)
      WRITE (6, 694) STDEV
  694 FORMAT (/T20,'THE STD. DEV. OF THE MAX K-S STATISTICS IS',F10.7)
  714 CONTINUE
      WRITE (6, 567)
  567 FORMAT (/T5,'**********************************************************')
C
C
      STOP
      END
```

```
      SUBROUTINE KSTEST (KTST, TMAX, NVAR)
C
CCCCCCCCCC
C        THIS SUBROUTINE PERFORMS A KOLMOGOROV-SMIRNOV TEST ON A SAMPLE
C        OF SIZE NVAR, AGAINST THE NORMAL(0,1) DISTRIBUTION
C        THE VALUE OF THE SAMPLE STEP FUNCTION, SN, IS EVALUATED AT EACH
C        POINT WHERE THE PROPOSED HYPOTHETICAL DISTRIBUTION FUNCTION, F,
C        TAKES ON VALUES OF MULTIPLES OF .01K(K = 1, 2, ... 100), AND THE
C        ABSOLUTE VALUE OF THE DIFFERENCE OF THE TWO FUNCTIONS,
C        ABS(F(X) - SN(X)) IS EVALUATED AT EACH SUCH POINT.  THE MAXIMUM
C        OF THE 100 DIFFERENCES IS RECORDED AS THE KOLMOGOROV-SMIRNOV
C        STATISTIC, D.
C
      DIMENSION RKTST1 (100), RKTST2 (100), IRSLT (100), RNORV (100)
      DIMENSION RKTST3 (36)
      COMMON RNORV
      DATA RKTST1/.01,.02,.03,.04,.05,.06,.07,.08,
     1 .09,.10,.11,.12,.13,.14,.15,.16,.17,.18,.19,.20,.21,.22,.23,.24,
     1 .25,.26,.27,.28,.29,.30,.31,.32,.33,.34,.35,.36,.37,.38,.39,.40,
     1 .41,.42,.43,.44,.45,.46,.47,.48,.49,.50,.51,.52,.53,.54,.55,.56,
     1.57,.58,.59,.60,.61,.62,.63,.64,.65,.66,.67,.68,.69,.70,.71,.72,
     1.73,.74,.75,.76,.77,.78,.79,.80,.81,.82,.83,.84,.85,.86,.87,.88,
     1.89,.90,.91,.92,.93,.94,.95,.96,.97,.98,.99,1.00/
      DATA RKTST2/-2.33,-2.054,-1.882,-1.751,-1.645,-1.556,-1.474,
     1-1.405,-1.341,-1.283,-1.224,-1.175,-1.124,-1.08,-1.037,-.995,
     1-.954,-.915,-.878,-.842,-.807,-.772,-.739,-.707,-.676,
     1 -.643,-.613,-.583,-.553,-.524,-.496,-.467,-.44,-.412,-.386,-.358,
     1  -.331,-.305,-.28,-.253,-.227,-.202,-.177,-.151,-.126,-.10,-.075,
     1  -.05,-.025,0.00,.025,.05,.075,.10,.126,.151,.177,.202,.227,
     1  .253,.28,.305,.331,.358,.386,.412,.440,.467,.496,.524,.553,
     1  .583,.613,.643,.676,.707,.739,.772,.807,.842,.878,.915,.954,
     1   .995,1.037,1.08,1.124,1.175,1.224,1.283,1.341,1.405,1.474,
     1  1.556,1.645,1.751,1.882,2.054,2.33,6.00/
C
      DATA RKTST3/.975,.842,.708,.624,.565,.521,.486,.457,.432,.410,
     1.391,.375,.361,.349,.338,.328,.318,.309,.301,.294,.289,.284,.279,
     1.274,.27,.264,.258,.252,.246,.24,.238,.236,.234,.232,.23,.14/
C
CCCCCCCCCC
C        RKTST1 ARE THE MULTIPLES OF .01 AT WHICH VALUES OF F(X) THE
C        DIFFERENCE BETWEEN THE SAMPLE STEP FUNCTION AND THE PROPOSED
C        CUMULATIVE DISTRIBUTION IS EVALUATED
C
C        RKTST2 IS THE LIST OF VALUES OF X AT WHICH F(X) IS EQUAL TO
C        .01K (K = 1, 2, ... 100) FOR THE NORMAL(0,1) DISTRIBUTION
C
```

```
C
C          RKTST3(I) IS CRITICAL VALUE OF D FOR A SAMPLE OF SIZE I AT .05
C          LEVEL OF SIGNIFICANCE.  IF D IS GREATER THAN RKTST3(I), THE
C          HYPOTHESIS THAT THE SAMPLE IS FROM A NORMAL (0,1) DISTRIBUTION
C          IS REJECTED.
      IF (NVAR .LE. 35) GO TO 109
      RKTST3 (36) = 1.36 / SQRT(1.0 * NVAR)
      NSAM = 36
      GO TO 116
  109 CONTINUE
      NSAM = NVAR
  116 CONTINUE
      DO 111 J = 1, 100
      ISUM = 0
      DO 110 I = 1, NVAR
      N = 0
      IF (RNORV (I) .LE. RKTST2 (J)) N = 1
      ISUM = ISUM + N
  110 CONTINUE
      IRSLT (J) = ISUM
  111 CONTINUE
      TMAX = 0.0
      DO 115 J = 1, 100
      CTMAX = ABS (RKTST1 (J) - IRSLT (J) / (1.0 * NVAR))
      IF (CTMAX .GE. TMAX) TMAX = CTMAX
  115 CONTINUE
      KTST = 1
      IF (TMAX .GT. RKTST3(NSAM)) KTST = 0
      RETURN
      END

      SUBROUTINE GAUSS (IX, S, AM, V)
C
C          THIS SUBROUTINE GENERATES UNIVARIATE NORMAL RANDOM VARIABLES
C
      A = 0.0
      DO 50 I = 1,12
      CALL RANDU (IX, IY, Y)
      IX = IY
   50 A = A + Y
      V = (A - 6.0) * S + AM
      RETURN
      END

       SUBROUTINE RANDU (IX, IY, YFL)
C          THIS SUBROUTINE GENERATES UNIFORM (0,1) RANDOM VARIABLES
```

```
      IY = IX * 65539
      IF (IY) 5,6,6
5     IY = IY + 2147483647 + 1
6     YFL = IY
      YFL = YFL * .4656613E-9
      RETURN
      END
```

BIBLIOGRAPHY

1. Ostle, B., Statistics in Research, Iow State, 1963.

2. Simpson, P. B., "Note on the Estimation of a Bivariate Distribution Function,: Annals of Mathematical Statistics, v. 22, p. 476-478, 1951.

3. Rosenblatt, M., "Remarks on a Multivariate Transformation," Annals of Mathematical Statistics, v. 22 p. 470.472, 1952.

4. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1960.

5. Kolmogorov, A., "Sulla Determinazione Empirica di Una Legge di Distribuzione," Giornale dell'Istitute degi Attuari, v. 4, p. 1-11, 1933.

6. Smirnov. H., "Sur les Ecarts de la Courbe de Distribution Empirique," Recueil Mathematique, v. 6, p. 3-26, 1939.

7. Massey, F. J., "The Kolmogorov-Smirnov Test for Goodness of Fit," Journal of the American Statistical Association, v. 46, p. 68-78, March 1951.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst. Professor D. R. Barr Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1
4.	LCOL James E. Miller, 067320, USMC Third Marine Aircraft Wing Marine Corps Air Station El Toro, California	1
5.	Commandant of the Marine Corps (Code AO3C) Headquarters, U.S. Marine Corps Washington, D.C. 20380	1
6.	James Carson Breckinridge Library Marine Corps Development and Educational Command Quantico, Virginia 22134	1
7.	Department of Operations Analysis (Code 55) Naval Postgraduate School Monterey, California 93940	1
8.	Commandant of the Marine Corps (Code AX) Headquarters, U.S. Marine Corps Washington, D.C., 20380	1

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

A GOODNESS OF FIT TEST FOR BIVARIATE NORMAL DISTRIBUTION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Master's Thesis; April 1970

5. AUTHOR(S) *(First name, middle initial, last name)*

James Edward Miller, Lieutenant Colonel, United States Marine Corps

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1970	40	7

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

This paper is an investigation of a goodness of fit test for bivariate normal distributions. The test procedure is based on random linear functions of bivariate normal random variables. The test makes use of the maximum Kolmogorov D(M) statistic over the linear functions which are computed. An estimate of the distribution of M is obtained by computer simulation. No attempt is made to determine the power of the test.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
GENERATION OF BIVARIATE NORMAL RANDOM VARIABLES						
GOODNESS OF FIT TEST						
BIVARIATE NORMAL DISTRIBUTION						

Thesis
M5873
c.1

119928

Miller
A goodness of fit
test for bivariate
normal distributions.